

Scaled Bregman divergences in a Tsallis scenario

R. C. Venkatesan ^{a,*} A. Plastino ^b

^a*Systems Research Corporation, Aundh, Pune 411007, India*

^b*IFLP, National University La Plata & National Research Council (CONICET)
C. C., 727 1900, La Plata, Argentina*

Abstract

There exist two different versions of the Kullback-Leibler divergence (K-Ld) in Tsallis statistics, namely the usual generalized K-Ld and the generalized Bregman K-Ld. Problems have been encountered in trying to reconcile them. A condition for consistency between these two generalized K-Ld-forms by recourse to the additive duality of Tsallis statistics is derived. It is also shown that the usual generalized K-Ld subjected to this additive duality, known as the dual generalized K-Ld, is a scaled Bregman divergence. This leads to an interesting conclusion: the dual generalized mutual information is a scaled Bregman information. The utility and implications of these results are discussed.

Key words: Generalized Tsallis statistics, additive duality, Kullback-Leibler divergence, scaled Bregman divergences, scaled Bregman information.

PACS: 05.20.-y; 89.70.-a

1 Introduction

The generalized statistics of Tsallis' has recently been the focus of much attention in statistical physics, complex systems, and allied disciplines (in this paper the terms generalized statistics, nonadditive statistics, and nonextensive statistics are indistinctly used)[1]. It is well-known that nonadditive statistics generalizes the extensive Boltzmann-Gibbs-Shannon (B-G-S) statistics. Its scope has lately been extended to studies of lossy data compression in

* Corresponding author.

Email addresses: ravi@systemsresearchcorp.com (R. C. Venkatesan),
plastino@venus.fisica.unlp.edu.ar (A. Plastino).

communication theory [2] and machine learning [3,4]. In this paper, attention is focussed upon the Tsallis-generalization of the concept of relative entropy, also known as Kullback-Leibler divergence (K-Ld), that constitutes a fundamental distance-measure in information theory [5]. The generalized K-Ld [6] encountered in deformed statistics has been described by Naudts [7] both as a special form of f-divergences [8, 9], and also in terms of Bregman-divergences [10]. Bregman divergences are, in turn, information geometric tools that have lately acquired great significance in a variety of disciplines ranging from information retrieval [11] and lossy data compression-machine learning [12] to statistical physics [13].

The generalized K-Ld is defined as [7]

$$D_\phi(p||r) = -\sum_i p_i \omega_\phi\left(\frac{r_i}{p_i}\right) = \frac{1}{\kappa} \sum_i p_i \left[\left(\frac{p_i}{r_i}\right)^\kappa - 1\right], \quad (1)$$

where p is an arbitrary distribution, r is the reference distribution, and κ is some nonadditivity parameter satisfying $-1 \leq \kappa \leq 1; \kappa \neq 0$. Here (1) employs the definition of the so-called *deduced logarithm* [7]

$$\omega_\phi(x) = \frac{1}{\kappa} (1 - x^{-\kappa}) . \quad (2)$$

An alternate form of the generalized K-Ld derived from the theory of Bregman divergences [7] is shown to be

$$D_\phi^B(p||r) = S_\phi(r) - S_\phi(p) - \sum_i (p_i - r_i) \ln_\phi(r_i), \quad (3)$$

where the generalized entropy and the deformed logarithm are defined as

$$S_\phi(z) = \sum_i z_i \omega_\phi\left(\frac{1}{z_i}\right), \quad (4)$$

and

$$\ln_\phi(x) = (1 + \kappa^{-1})(x^\kappa - 1), \quad (5)$$

respectively.

1.1 Problems reconciling the Tsallis versions of the Kullback-Leibler divergence

Specializing the above concepts to the Tsallis scenario by setting $\kappa = q - 1$, Eqs. (1) and (3) yield the usual doubly convex generalized K-Ld [6]

$$D_{K-L}^q(p||r) = \frac{1}{q-1} \sum_i p_i \left[\left(\frac{p_i}{r_i} \right)^{q-1} - 1 \right], \quad (6)$$

and the generalized Bregman K-Ld

$$D_q^B[p||r] = \frac{1}{(q-1)} \sum_i p_i (p_i^{q-1} - r_i^{q-1}) - \sum_i (p_i - r_i) r_i^{q-1}, \quad (7)$$

respectively.

While the form of the generalized Bregman K-Ld (BK-Ld) is more appealing than (6) from an information geometric viewpoint, it does contain certain inherent drawbacks.

A study by Abe and Bagci [13] has demonstrated that the generalized K-Ld defined by (6) is jointly convex in terms of both p_i and r_i while the form defined by (7) is convex *only* in terms of p_i . A further distinction between the two forms of the generalized K-Ld concerns the property of composability. While the form defined by (6) is composable, the form defined by (7) does not exhibit this property. The fact that the two generalized K-Ld versions have no apparent relation to each other should be a cause of concern for practitioners of nonextensive statistical physics.

A second issue to address concerns the manner in which mean values are computed. Nonextensive statistics has employed a number of forms in which expectations may be defined. Prominent among these are the linear constraints originally employed by Tsallis [1] (also known as normal averages) of the form: $\langle A \rangle = \sum_i p_i A_i$, the Curado-Tsallis (C-T) constraints [14] of the form: $\langle A \rangle_q = \sum_i p_i^q A_i$, and the normalized Tsallis-Mendes-Plastino (TMP) constraints [15] (also known as q -averages) of the form: $\langle \langle A \rangle \rangle_q = \sum_i \frac{p_i^q}{\sum_i p_i^q} A_i$.

A fourth constraining procedure is the optimal Lagrange multiplier (OLM) approach [16]. Of these four methods to describe expectations, the most commonly employed by Tsallis-practitioners is the TMP-one.

Recent works by Abe [17, 18] suggest that in generalized statistics expectations defined in terms of normal averages, in contrast to those defined by q -averages, are consistent with the generalized H-theorem and the generalized *Stosszahlansatz* (molecular chaos hypothesis). The correctness of normal average expectations vis-à-vis q -average (or TMP) ones has also been investigated by Hasegawa [19, 20]. Understandably, a re-formulation of the variational perturbation approximations in nonextensive statistical physics followed [21], via an application of q -deformed calculus [22].

Further concern is originated by a consistency issue. This stems from the

fact that the form of the generalized K-Ld defined by (6) is consistent with expectations and constraints defined by q -averages while, on the other hand, the generalized Bregman K-Ld defined by (7) is consistent with expectations defined by normal averages [13].

1.2 Additive duality

The additive duality is a fundamental property in generalized statistics. One implication of the additive duality is that it permits a deformed logarithm defined by a given nonadditivity parameter (say, q) to be inferred from its *dual deformed* logarithm [1, 2, 23] parameterized by: $q^* = 2 - q$.

Our leitmotif for invoking the additive duality stems from the form of the BK-Ld (7). Setting $\kappa = q - 1$ in (2) and (5) yields a Tsallis entropy of the form: $S_q(z) = -z \ln_q(z)$, which is the Tsallis entropy defined in Section 2.1 of this paper subjected to the re-parameterization $q \rightarrow 2 - q$.¹ Thus, in the Tsallis scenario, (5) is actually the dual Tsallis entropy defined in (15) with the additive duality ($q \rightarrow 2 - q$) *implicitly* accounted for. Given these facts, from the definition of Bregman divergences provided by Definition 1 in Section 2.3 below, the form of the BK-Ld (7) can only be obtained by specifying the complex generating function as: $\phi(z) = z \ln_q z$, followed by the re-parameterization $q \rightarrow 2 - q$. More specifically, the BK-Ld (7) can only be derived from first principles using (5) defined in the Tsallis scenario by recourse to the additive duality. Hence, the necessity for invoking the additive duality in this paper, where the re-parameterization is *explicitly* accounted for by defining: $q^* = 2 - q$.

By definition (see Section 2.1 below for details), the generalized K-Ld subjected to the additive duality is referred to as the dual generalized K-Ld having the form

$$D_{K-L}^{q^*}[p||r] = \sum_i p_i \ln_{q^*} \left(\frac{p_i}{r_i} \right) = \frac{1}{(1 - q^*)} \sum_i (p_i^{2-q^*} r_i^{q^*-1} - 1). \quad (8)$$

However, employing the definitions of Bregman divergences presented in Section 2.3 below, the BK-Ld is of the form

$$D_{q^*}^B[p||r] = \frac{1}{(1 - q^*)} \sum_i p_i (p_i^{1-q^*} - r_i^{1-q^*}) - \sum_i (p_i - r_i) r_i^{1-q^*}, \quad (9)$$

for the convex generating function: $\phi(z) = z \ln_{q^*} z$.

¹ Here " \rightarrow " denotes a re-parameterization of the nonadditivity parameter, and is not a limit.

1.3 Goal of this paper

Scaled Bregman divergences, formally introduced by Stummer [24] and Stummer and Vajda [25], unify separable Bregman divergences [10] (defined below in Section 2.3) and f-divergences [8,9]. This paper uses scaled Bregman divergences as its basis, and accomplishes the following objectives:

- (i) the generalized K-Ld defined by (6) subjected to the additive duality (dual generalized K-Ld (8) and (15)) is shown to be consistent with the canonical probability that maximizes the dual Tsallis entropy of the form [2, 26]: $S_{q^*} = -\sum_i p_i \ln_{q^*} p_i$ employed in conjunction with expectations defined by normal averages (Section 3 of this paper),
- (ii) a correspondence between the dual generalized K-Ld and the generalized Bregman K-Ld is derived (Section 4 below),
- (iii) the dual generalized K-Ld is demonstrated to be a scaled Bregman divergence and that its expectation is a *scaled Bregman information*, i.e. the expectation of a scaled Bregman divergence (Section 5 below) for both regimes of the dual nonadditivity parameter $0 < q^* < 1$ and $q^* > 1$ [27] (Section 5 below).

Section 6 is devoted to discussion and conclusions. The primary conclusion of this paper is the necessity of employing the dual generalized K-Ld when performing a minimum cross entropy analysis (principle of minimum discrimination information) of Kullback [28] and Kullback and Khairat [29] using constraints defined by *normal average expectations*.

2 Theoretical preliminaries

The essential concepts around which this communication revolves are reviewed in the three subsections that follow.

2.1 Tsallis entropy and the additive duality

By definition, the Tsallis entropy, is defined in terms of discrete variables as [1]

$$S_q(X) = -\frac{1 - \sum_x p^q(x)}{1-q}; \sum_x p(x) = 1. \quad (10)$$

The constant q is referred to as the nonadditive parameter. Here, (10) implies that extensive B-G-S statistics is recovered as $q \rightarrow 1$. Taking the limit $q \rightarrow 1$ in (10) and invoking l'Hospital's rule, $S_q(X) \rightarrow S(X)$, i.e., the Shannon

entropy. Nonextensive statistics is intimately related to *q-deformed* algebra and calculus (see [22] and the references within). The *q-deformed* logarithm and exponential are defined as [22]

$$\begin{aligned} \ln_q(x) &= \frac{x^{1-q}-1}{1-q}, \\ \text{and,} \\ \exp_q(x) &= \begin{cases} [1 + (1-q)x]^{\frac{1}{1-q}}; & 1 + (1-q)x \geq 0 \\ 0; & \text{otherwise,} \end{cases} \end{aligned} \quad (11)$$

respectively. In this respect, an important relation from *q-deformed* algebra is [2, 22, 27]

$$\ln_q\left(\frac{x}{y}\right) = y^{q-1}(\ln_q x - \ln_q y). \quad (12)$$

The Tsallis entropy (10), conditional Tsallis entropy, and, joint Tsallis entropy may be written as [1]

$$\begin{aligned} S_q(X) &= -\sum_x p(x)^q \ln_q p(x), \\ S_q(\tilde{X}|X) &= -\sum_x \sum_{\tilde{x}} p(x, \tilde{x})^q \ln_q p(\tilde{x}|x), \\ S_q(X, \tilde{X}) &= -\sum_x \sum_{\tilde{x}} p(x, \tilde{x})^q \ln_q p(x, \tilde{x}) \\ &= S_q(X) + S_q(\tilde{X}|X) = S_q(\tilde{X}) + S_q(X|\tilde{X}), \end{aligned} \quad (13)$$

respectively.

This paper makes prominent use of the *additive duality* in nonextensive statistics. Setting $q^* = 2-q$, from (11) the *dual deformed* logarithm and exponential are defined as

$$\ln_{q^*}(x) = -\ln_q\left(\frac{1}{x}\right), \text{ and, } \exp_{q^*}(x) = \frac{1}{\exp_q(-x)}. \quad (14)$$

The dual Tsallis entropy, the dual conditional Tsallis entropy, the dual joint Tsallis entropy, and, the dual generalized K-Ld may thus be written as

$$\begin{aligned} S_{q^*}(X) &= -\sum_x p(x) \ln_{q^*} p(x), \\ S_{q^*}(\tilde{X}|X) &= -\sum_x \sum_{\tilde{x}} p(x, \tilde{x}) \ln_{q^*} p(\tilde{x}|x), \\ S_{q^*}(X, \tilde{X}) &= -\sum_x \sum_{\tilde{x}} p(x, \tilde{x}) \ln_{q^*} p(x, \tilde{x}) \\ &= S_{q^*}(X) + S_{q^*}(\tilde{X}|X) = S_{q^*}(\tilde{X}) + S_{q^*}(X|\tilde{X}), \end{aligned} \quad (15)$$

and,

$$D_{K-L}^{q^*}[p(X) \| r(X)] = \sum_x p(x) \ln_{q^*} \frac{p(x)}{r(x)},$$

respectively. The dual Tsallis entropy has already been studied in a maximum (Tsallis) entropy setting (for example, see Ref. [26]). Note that the dual Tsallis entropy acquires a form identical to the B-G-S entropies, with $\ln_{q^*}(\bullet)$ replacing $\log(\bullet)$. It is important to note that the $q^* = 2 - q$ duality has been studied within the Sharma-Taneja-Mittal framework by Kaniadakis, *et. al.* [30]. The dual Tsallis entropy has been demonstrated to support a parametrically extended information theory, as is defined in Theorem 2 below.

Theorem 1 [2]: *Let $X_1, X_2, X_3, \dots, X_n$ be random variables obeying the probability distribution $p(x_1, x_2, x_3, \dots, x_n)$, then we have the chain rule*

$$S_{q^*}(X_1, X_2, X_3, \dots, X_n) = \sum_{i=1}^n S_{q^*}(X_i | X_{i-1}, \dots, X_1). \quad (16)$$

2.2 Generalized mutual informations

Given a random variable X in \mathcal{X} where instances of X are $x_1, \dots, x_{|\mathcal{X}|}$, for $0 < q < 1$, the generalized mutual information is defined in terms of the generalized K-Ld [2]

$$I_{0 < q < 1}(X; \tilde{X}) = - \sum_{x, \tilde{x}} p(x, \tilde{x}) \ln_q \left(\frac{p(x) p(\tilde{x})}{p(x, \tilde{x})} \right). \quad (17)$$

For nonadditivity parameters in the range $q > 1$, the generalized mutual information is [2,27]

$$\begin{aligned} I_q(X; \tilde{X}) &= S_q(X) - S_q(X | \tilde{X}) = S_q(\tilde{X}) - S_q(\tilde{X} | X) \\ &= S_q(X) + S_q(\tilde{X}) - S_q(X, \tilde{X}) = I_q(\tilde{X}; X); q > 1 \end{aligned} \quad (18)$$

For (18) to hold true, the inequalities (*sub-additivities*)

$$S_q(X | \tilde{X}) \leq S_q(X), \text{ and, } S_q(\tilde{X} | X) \leq S_q(\tilde{X}), \quad (19)$$

have to hold true. This is not guaranteed for nonadditivity parameters in the range $0 < q < 1$ [2,27].

As stated in Refs. [2] and [27], the generalized mutual information is separately defined within two separate q -ranges $0 < q < 1$ and $q > 1$. They have different uses. For $0 < q < 1$, the generalized mutual information, as defined by (17), provides a means of extrapolating the Csiszár-Tusndy theory [31] to the

nonextensive domain for two convex sets of probability distributions [2]. This has important implications in communication theory and allied disciplines [2, 5].

For $q > 1$, the generalized mutual information as defined by (18) possesses a number of important properties such as the *generalized data processing inequality* and the *generalized Fano inequality* [27]. This allows one to define Lagrangians and cost functions for processes defined by a Markov chain relation.

Theorem 2 [2] The generalized mutual information for nonadditivity parameters in the range $0 < q < 1$ and $q > 1$ are related via the additive duality

$$I_{q^*}(X; \tilde{X}) = - \sum_x \sum_{\tilde{x}} p(x, \tilde{x}) \ln_{q^*} \left(\frac{p(x)p(\tilde{x})}{p(x, \tilde{x})} \right) \quad (20)$$

$$\stackrel{(q^* \rightarrow q)}{=} S_q(X) + S_q(\tilde{X}) - S_q(X, \tilde{X}) = I_q(X; \tilde{X}); 0 < q^* < 1, \text{ and } q > 1.$$

2.3 Bregman divergences and scaled Bregman divergences

This sub-section introduces the formal definition of Bregman divergences and some of their select properties. The Bregman divergence or Bregman distance is similar to a metric, but does not in general satisfy the triangle inequality nor symmetry. Bregman divergences do however obey the Pythagorean theorem (for example, see Appendix A in [12]). There are two ways in which Bregman divergences are important. Firstly, they generalize squared Euclidean distances to a class of distances that all share similar properties. Secondly, they bear a strong connection to exponential families of distributions. There is a bijection between regular exponential families and regular Bregman divergences. Bregman divergences are named after L. M. Bregman [10], who introduced the concept in 1967. More recently researchers in geometric algorithms have shown that many important algorithms can be generalized from Euclidean metrics to distances defined by Bregman divergence. This sub-section introduces the formal definition of Bregman divergences and some of their properties.

Definition 1 (Bregman divergences)[10, 32]: Let ϕ be a real valued strictly convex function defined on the convex set $\mathcal{S} \subseteq \text{dom}(\phi)$, the domain of ϕ such that ϕ is differentiable on $\text{ri}(\mathcal{S})$, the relative interior of \mathcal{S} . The Bregman divergence $B_\phi : \mathcal{S} \times \text{ri}(\mathcal{S}) \mapsto [0, \infty)$ is defined as: $B_\phi(z_1, z_2) = \phi(z_1) - \phi(z_2) - \langle z_1 - z_2, \nabla \phi(z_2) \rangle$, where: $\nabla \phi(z_2)$ is the gradient of ϕ evaluated at z_2 .²

² Note that $\langle \bullet, \bullet \rangle$ denotes the inner product. Calligraphic fonts denote sets.

Definition 2 (Notations)[25]: \mathcal{M} denotes the space of all finite measures on a measurable space $(\mathcal{X}, \mathcal{A})$ and $\mathcal{P} \subset \mathcal{M}$ the subspace of all probability measures. Unless otherwise explicitly stated P, R, M are mutually measure-theoretically equivalent measures on $(\mathcal{X}, \mathcal{A})$ dominated by a σ -finite measure λ on $(\mathcal{X}, \mathcal{A})$. Then the densities

$$p = \frac{dP}{d\lambda}, r = \frac{dR}{d\lambda}, \text{ and } m = \frac{dM}{d\lambda}, \quad (21)$$

have a common support which will be identified with \mathcal{X} . Unless stated otherwise, it is assumed that $P, R \in \mathcal{P}, M \in \mathcal{M}$ and that $\phi : (0, \infty) \mapsto \mathcal{R}$ is a continuous and convex function.

Definition 3 (Scaled Bregman Divergences) [25] *The Bregman divergence of probability measures P, R scaled by an arbitrary measure M on $(\mathcal{X}, \mathcal{A})$ measure-theoretically equivalent with P, R is defined by*

$$\begin{aligned} B_\phi(P, R | M) &= \int_{\mathcal{X}} \left[\phi\left(\frac{p}{m}\right) - \phi\left(\frac{r}{m}\right) - \left(\frac{p}{m} - \frac{r}{m}\right) \nabla \phi\left(\frac{r}{m}\right) \right] dM \\ &= \int_{\mathcal{X}} \left[m\phi\left(\frac{p}{m}\right) - m\phi\left(\frac{r}{m}\right) - (p - r) \nabla \phi\left(\frac{r}{m}\right) \right] d\lambda. \end{aligned} \quad (22)$$

The convex ϕ may be interpreted as the generating function of the divergence. In a discrete setting, a scaled Bregman divergence is defined as [25]

$$B_\phi(p, r | m) = \sum_{i=1}^d \left[\phi\left(\frac{p_i}{m_i}\right) - \phi\left(\frac{r_i}{m_i}\right) - \left(\frac{p_i}{m_i} - \frac{r_i}{m_i}\right) \nabla \phi\left(\frac{r_i}{m_i}\right) \right] m_i. \quad (23)$$

3 Maximum dual Tsallis entropy models

The Tsallis entropy parameterized by q is defined as [1]

$$S_q[p] = -\frac{1 - \sum_i p_i^q}{(1 - q)}. \quad (24)$$

Setting $q^* = 2 - q$, the dual Tsallis entropy is expressed as [2, 26]

$$S_{q^*}[p] = -\frac{1 - \sum_i p_i^{2-q^*}}{(q^* - 1)}. \quad (25)$$

The q^* -deformed Lagrangian (normal averages used) to be extremized reads

$$\Phi_*[p, \alpha, \beta] = S_{q^*}[p] - \alpha \left(\sum_i p_i - 1 \right) - \beta \left(\sum_i p_i E - U \right), \quad (26)$$

yielding the canonical probability that maximizes the dual Tsallis entropy as

$$p_i = \frac{\left[1 - \frac{(1-q^*)}{\aleph_{q^*}} \tilde{\beta}^*(E-U)\right]^{\frac{1}{1-q^*}}}{\aleph_{q^*}^{\frac{1}{q^*-1}}}, \quad (27)$$

$$\aleph_{q^*} = \sum_i p_i^{2-q^*} = \tilde{Z}(\tilde{\beta}^*)^{q^*-1}, \text{ and, } \tilde{\beta}^* = \frac{\beta}{2-q^*}$$

Note that the methodology developed in [33] is employed in the maximum Tsallis analysis using constraints defined by normal averages. Here, $\tilde{Z}(\tilde{\beta}^*)$ is the canonical partition function. The Appendix in this paper provides the detailed derivation of (27). The dual Tsallis entropy is defined as

$$S_{q^*}[p] = \frac{\aleph_{q^*} - 1}{(q^* - 1)}. \quad (28)$$

Here, $\tilde{\beta}^* = \beta/(2-q^*)$ is referred to as the "dual scaled inverse thermodynamic temperature".

4 Correspondence between the generalized Kullback-Leibler divergences

The generalized free energy (GFE) for normal averages expectations is defined as [21]

$$F_q = U - \frac{1}{\tilde{\beta}} S_q[p]. \quad (29)$$

Note that the expression for the GFE (29) has recently been the object of much research and debate. The effective inverse temperature $\tilde{\beta}$ is the energy Lagrange multiplier scaled with respect to q . The energy Lagrange multiplier generally relates to the thermodynamic temperature T as: $\beta = \frac{1}{k_B T}$, where k_B is the Boltzmann constant (sometimes set to unity for the sake of convenience) *only in the limiting case* $q \rightarrow 1$. Prominent attempts to clarify this issue are those by Abe *et. al.* [34], Abe [35], amongst others. Similarly, the q^* -deformed (dual) GFE is defined as

$$F_{q^*} = U - \frac{1}{\tilde{\beta}^*} S_{q^*}[p]. \quad (30)$$

At this stage, setting the reference probability $r = \tilde{p}$ in (9), and associating the quantities \tilde{S}_q, \tilde{U} , and, $\tilde{\aleph}_q$ with the maximum Tsallis entropy canonical distribution \tilde{p}_i , yields

$$D_{q^*}^B[p||\tilde{p}] = \frac{1}{(1-q^*)} \sum_i p_i \left(p_i^{1-q^*} - \tilde{p}_i^{1-q^*} \right) - \sum_i (p_i - \tilde{p}_i) \tilde{p}_i^{1-q^*}. \quad (31)$$

Substituting (27) into (31) one gets

$$\begin{aligned} D_{q^*}^B [p \parallel \tilde{p}] &= \frac{1}{(1-q^*)} \sum_i p_i \left(\aleph_{q^*} - (1-q^*) \tilde{\beta}^* U - \tilde{\aleph}_{q^*} + (1-q^*) \tilde{\beta}^* \tilde{U} \right) \\ &\quad - \sum_i (p_i - \tilde{p}_i) \left(\tilde{\aleph}_{q^*} + (1-q^*) \tilde{\beta}^* (E - U) \right). \end{aligned} \quad (32)$$

With the aid of (28) and (30) and the normalization property, (32) leads now to

$$D_{q^*}^B [p_i \parallel \tilde{p}_i] = \tilde{\beta}^* [F_{q^*} - \tilde{F}_{q^*}], \quad (33)$$

and the dual generalized K-Ld defined in (15) becomes

$$\begin{aligned} D_{K-L}^{q^*} [p_i \parallel \tilde{p}_i] &= \sum_i p_i \ln_{q^*} \left(\frac{p_i}{\tilde{p}_i} \right) \\ &= \frac{1}{(1-q^*)} \sum_i p_i \left(\frac{p_i^{1-q^*}}{\tilde{p}_i^{1-q^*}} - 1 \right) = \frac{1}{(1-q^*)} \sum_i p_i \left(\frac{p_i^{1-q^*} - \tilde{p}_i^{1-q^*}}{\tilde{p}_i^{1-q^*}} \right) \\ &= \beta^* [F_{q^*} - \tilde{F}_{q^*}] \tilde{\Psi}_{q^*} \sum_i p_i = \beta^* [F_{q^*} - \tilde{F}_{q^*}] \tilde{\Psi}_{q^*}; \tilde{\Psi}_{q^*} = \sum_i \tilde{p}_i^{q^*-1}. \end{aligned} \quad (34)$$

From (33) and (34), the correspondence relation between the usual generalized K-Ld, the dual generalized K-Ld, and the generalized Bregman K-Ld is

$$\begin{aligned} D_{K-L}^B [p_i \parallel \tilde{p}_i] &= \frac{D_{K-L}^{q^*} [p_i \parallel \tilde{p}_i]}{\tilde{\Psi}_{q^*}} \stackrel{q^* \rightarrow q}{=} \frac{D_{K-L}^q [p_i \parallel \tilde{p}_i]}{\tilde{\Psi}_q}, \\ \tilde{\Psi}_q &= \sum_i \tilde{p}_i^{1-q}, \end{aligned} \quad (35)$$

which is a compact result.

It is important to point out that one application of the correspondence relation presented in this Section is that of providing an alternate means to derive the dual generalized K-Ld from the generalized Bregman K-Ld. This may be accomplished by invoking the linearity property of Bregman divergences (see Appendix A of Ref. [12]).

The above mentioned linearity property states that the Bregman divergence is a linear operator i.e., $\forall x \in \mathcal{S}, y \subset ri(S)$ (where $ri(\bullet)$ denotes the relative interior of a set), $B_{c\phi}(x, y) = cB_\phi(x, y)$ (for $c > 0$). From (35), it is immediately evident that multiplying (31) by: $c = \tilde{\Psi}_{q^*} > 0$ and invoking (12) readily yields the dual generalized K-Ld. This relation between the dual generalized K-Ld and Bregman divergences may however be viewed as one of convenience, which although tenable, lacks the formal theoretical rigor of the results presented in Section 5 below.

5 Dual generalized K-Ld, scaled Bregman divergences, and the scaled Bregman information

This Section serves a two-fold purpose: (i) it is established that the dual generalized K-Ld defined in (15) is a scaled Bregman divergence, (ii) we introduce the concept of scaled Bregman information as the expectation of a scaled Bregman divergence.

5.1 Dual generalized K-Ld as a scaled Bregman divergence

Let (i) $t = \frac{z}{m}$ and (ii) the generating function of the Bregman divergence be a convex function $\phi(t)$, with m the scaling. For a generating function $\phi(t) = t \ln_{q^*} t$, the discrete form of the scaled Bregman divergence (23) acquires the form

$$\begin{aligned} B_\phi(p, r | m) &= \sum_i \left[\frac{p_i}{m_i} \ln_{q^*} \frac{p_i}{m_i} - \frac{r_i}{m_i} \ln_{q^*} \frac{r_i}{m_i} - \left(\frac{p_i}{m_i} - \frac{r_i}{m_i} \right) \nabla \frac{r_i}{m_i} \ln_{q^*} \left(\frac{r_i}{m_i} \right) \right] m_i. \\ &= \sum_i \left[p_i \ln_{q^*} \frac{p_i}{m_i} - p_i \ln_{q^*} \frac{r_i}{m_i} - (p_i - r_i) \left(\frac{r_i}{m_i} \right)^{1-q^*} \right] \\ &= \sum_i \left\{ p_i m_i^{q^*-1} [\ln_{q^*} p_i - \ln_{q^*} r_i] - (p_i - r_i) \left(\frac{r_i}{m_i} \right)^{1-q^*} \right\}. \end{aligned} \quad (36)$$

At this point, specifying $m_i = r_i$ in (36), and invoking (12) and the normalization relation: $\sum_i p_i = \sum_i r_i = 1$, the dual generalized K-Ld in (15) is recovered, i.e.

$$B_\phi(p, r | m = r) = \sum_i p_i \ln_{q^*} \left(\frac{p_i}{r_i} \right). \quad (37)$$

This is a q^* -deformed f-divergence and is consistent with the theory derived in Refs. [24] and [25], when extended to deformed statistics. The above result may also be employed in the case of the dual generalized K-Ld between a conditional probability and a marginal probability. Let X and Y be random variables in \mathcal{X} and \mathcal{Y} respectively. Let the marginal discrete probability measures be: $\{p(x_i)\}_{i=1}^n$ and $\{p(y_j)\}_{j=1}^m$, respectively. In such circumstances, the dual generalized K-Ld reads

$$D_{K-L}^{q^*} [p(Y | x_i) || p(Y)] = \sum_j p(y_j | x_i) \ln_{q^*} \frac{p(y_j | x_i)}{p(y_j)}, \quad (38)$$

and is indeed a scaled Bregman divergence with the scaling: $p(y_j)$.

5.2 Dual generalized K-Ld and the scaled Bregman information

Definition 4 [36]: For any Bregman divergence (or scaled Bregman divergence) $B_\phi : \mathcal{S} \times \text{int}(\mathcal{S}) \mapsto \mathbb{R}^+$ and any random variable $Z \sim w(z)$ (where $w(z)$ is the probability measure associated with Z), $z \in \mathcal{Z} \subseteq \mathcal{S}$, **the Bregman information** (or **scaled Bregman information**) which is a measure of the information in Z is defined as

$$I_\phi(Z) = \langle B_\phi(Z, \langle Z \rangle) \rangle. \quad (39)$$

For example, let X be a random variable that takes values in $\mathcal{X} = \{x_i\}_{i=1}^n$ following a probability measure $p(\mathbf{x})$. Let $\mu = \langle X \rangle = \sum_i p(x_i)x_i$, and let B_ϕ be a Bregman divergence (or scaled Bregman divergence). Then the Bregman information (or scaled Bregman information) of X is defined as

$$I_\phi(X) = \sum_i p(x_i) B_\phi(x_i, \mu). \quad (40)$$

Consider a random variable Z_x which takes values in the set of probability distributions: $\mathcal{Z}_x = \{p(Y|x_i)\}_{i=1}^n$, following the marginal probability: $\{p(x_i)\}_{i=1}^n$ defined over this set. The expectation of Z_x is

$$\mu = \sum_i p(x_i) p(Y|x_i) = \sum_i p(x_i, Y) = p(Y). \quad (41)$$

Thus, from (38)-(41) the scaled Bregman information, which is the dual generalized mutual information, may be defined as

$$\begin{aligned} I_{q^*}(X; Y) &= \sum_i p(x_i) D_{K-L}^{q^*}[p(Y|x_i) \| p(Y)] \\ &= \sum_i p(x_i) \sum_j p(y_j|x_i) \ln_{q^*} \frac{p(y_j|x_i)}{p(y_j)} \\ &= D_{K-L}^{q^*}[p(X, Y) \| p(X)p(Y)] = I_\phi(Z_x). \end{aligned} \quad (42)$$

Similarly, the relation: $I_{q^*}(X; Y) = I_\phi(Z_y)$ also holds true, when Z_y is a random variable which takes values in the set of probability distributions: $\mathcal{Z}_y = \{p(X|y_j)\}_{j=1}^m$, following the marginal probability: $\{p(y_j)\}_{j=1}^m$ defined over this set. In this case, the scaled Bregman divergence is: $D_{K-L}^{q^*}[p(X|y_j) \| p(X)] = \sum_j p(y_j) \ln_{q^*} \frac{p(X|y_j)}{p(X)}$, and the normal averages expectation is calculated with respect to: $p(y_j)$.

For values: $q^* > 1$, the scaled Bregman information acquires the form

$$\begin{aligned}
I_\phi(Z_x) &= \sum_{i,j} p(x_i, y_j) \ln_{q^*} \frac{p(y_j|x_i)}{p(y_j)} \\
&\stackrel{(a)}{=} \sum_{i,j} p(x_i, y_j) p(y_j)^{q^*-1} (\ln_{q^*} p(y_j|x_i) - \ln_{q^*} p(y_j)) \\
&\stackrel{(b)}{=} \sum_{i,j} p(x_i, y_j) \sum_i p(x_i)^{q^*-1} p(y_j|x_i)^{q^*-1} (\ln_{q^*} p(y_j|x_i) - \ln_{q^*} p(y_j)) \\
&= \sum_i \sum_j p(x_i) p(y_j|x_i) \sum_i p(x_i)^{q^*-1} p(y_j|x_i)^{q^*-1} (\ln_{q^*} p(y_j|x_i) - \ln_{q^*} p(y_j)) \\
&= \sum_i \sum_j p(x_i, y_j)^{q^*} (\ln_{q^*} p(y_j|x_i) - \ln_{q^*} p(y_j)) \\
&= - \sum_j p(y_j)^{q^*} \ln_{q^*} p(y_j) + \sum_{i,j} p(x_i, y_j)^{q^*} \ln_{q^*} p(y_j|x_i) \\
&= - \sum_i p(x_i)^{q^*} \ln_{q^*} p(x_i) + \sum_{i,j} p(x_i, y_j)^{q^*} \ln_{q^*} p(x_i|y_j).
\end{aligned} \tag{43}$$

In the derivation (43) (a) denotes the use of (12) while (b) denotes setting: $p(y_j)^{q^*-1} = \sum_i p(x_i)^{q^*-1} p(y_j|x_i)^{q^*-1}$. Note that the last two expressions in (43) are identical to (18) with the nonadditivity parameter q^* replacing q in (18). Defining

$$\begin{aligned}
\tilde{S}_{q^*}(X) &= - \sum_x p(x)^{q^*} \ln_{q^*} p(x), \\
\tilde{S}_{q^*}(\tilde{X}|X) &= - \sum_x \sum_{\tilde{x}} p(x, \tilde{x})^{q^*} \ln_{q^*} p(\tilde{x}|x), \\
\tilde{S}_{q^*}(X, \tilde{X}) &= - \sum_x \sum_{\tilde{x}} p(x, \tilde{x})^{q^*} \ln_{q^*} p(x, \tilde{x}) \\
&= \tilde{S}_{q^*}(X) + \tilde{S}_{q^*}(\tilde{X}|X) = \tilde{S}_{q^*}(\tilde{X}) + \tilde{S}_{q^*}(X|\tilde{X}),
\end{aligned} \tag{44}$$

the scaled Bregman information (43) acquires the form

$$I_\phi(Z_x) = \tilde{S}_{q^*}(Y) - \tilde{S}_{q^*}(Y|X) = \tilde{S}_{q^*}(X) - \tilde{S}_{q^*}(X|Y) = I_\phi(Z_y), \tag{45}$$

where the inequalities: $\tilde{S}_{q^*}(Y|X) \leq \tilde{S}_{q^*}(Y)$, and, $\tilde{S}_{q^*}(X|Y) \leq \tilde{S}_{q^*}(X)$ hold true.

Comparison of (13) and (44) readily reveals that the original expressions for the Tsallis entropy and conditional Tsallis entropy, and their equivalent forms derived from the dual generalized mutual information (43), are invariant under interchange of the nonadditivity parameters q and $q^* = 2 - q$. While this is indeed an appealing observation, two points need to be noted: (i) the physics of the problem is defined by q and not q^* , and, (ii) Eqs. (13) and (44) correspond to two separate physical conditions.

The generalized mutual information (18) is expressed in terms of (13) for $q > 1$. This corresponds to probability distributions of particular interest to Tsallis statistics, i.e. "long-tailed" and power law distributions, amongst others. On the other hand, (43)-(45) correspond to $q^* > 1 \Rightarrow q < 1$. This regime is not of great interest in generalized statistics. Thus, when modeling problems in generalized statistics (for example, see Ref. [3]) whose variational principle requires invoking the properties Bregman divergences, use of Theorem 2 (Eq. (20)) is to be employed in order to simultaneously achieve information-geometric and physical consistency.

6 Summary and Discussions

Our present endeavors have enabled us to reach several findings regarding the Tsallis environment.

- The dual generalized K-Ld was shown to be a scaled Bregman divergence.
- With regards to expectation values computed using normal averages, the dual generalized mutual information was demonstrated to be a scaled Bregman information,
- The correspondence linking the dual generalized K-Ld, the generalized Bregman K-Ld (for probability distributions which maximize the dual Tsallis entropy when using normal-averages-constraints), and the usual form of the generalized K-Ld, has been established. Such a correspondence has not been previously investigated in Tsallis statistics literature.

From the analyses in Sections 3-5, it becomes obvious from a combined statistical physics plus information geometric perspective that the dual generalized K-Ld should also be employed as the measure of uncertainty when performing a minimum cross entropy analysis (principle of minimum discrimination information) [28, 29, 37] for constraints that employ normal averages.

A simpler justification stems from the fact that while in the orthodox B-G-S theory the K-Ld is a Bregman divergence [12], its Tsallis counterpart is not a Bregman divergence. Instead, as established in this paper, the dual generalized K-Ld is a scaled Bregman divergence. Future work uses the results derived herein to analyze: (i) the generalized statistics rate distortion theory [2], (ii) the generalized statistics information bottleneck method [3] within the context of scaled Bregman divergences and scaled Bregman informations, and (iii) deformed statistics extensions of the minimum Bregman information principle and their applications in machine learning [36].

Acknowledgements

RCV gratefully acknowledges support from *RAND-MSR* contract *CSM-DI & S-QIT-101155-03-2009*.

References

- [1] C. Tsallis, Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World, Springer, Berlin, 2009; C. Tsallis, J. Stat. Phys., **542** (1988) 479.
- [2] R. C. Venkatesan and A. Plastino, Physica A, **388**, 12 (2009) 2337.
- [3] R. C. Venkatesan and A. Plastino, "Deformed Statistics Formulation of the Information Bottleneck Method", Proceedings of the IEEE Int. Symp. on Information Theory 2009, 1323, 2009.
- [4] A. F. T. Martins, N. A. Smith, E. P. Xing, P. M. Q. Aguiar, and M. A. T. Figueiredo, J. Machine Learning Research (JMLR), **10** (2009) 935.
- [5] T. Cover and J. Thomas, Elements of Information Theory, John Wiley & Sons (New York, NY, 1991).
- [6] L. Borland, A. Plastino, and C. Tsallis, J. Math. Phys., **39** (1998) 6490.
- [7] J. Naudts, Rev. Math. Phys., **16** (2004) 809.
- [8] I. Csiszár, Studia Sci. Math. Hungar., **2** (1967) 299.
- [9] F. Liese and I. Vajda, Convex Statistical Distances, Teubner, Leipzig, 1987.
- [10] L. M. Bregman, USSR Computational Mathematics and Mathematical Physics, **7** (1967) 200.
- [11] M.N. Do and M. Vetterli, IEEE Trans. Image Proc., **11** (2002) 146.
- [12] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, J. Machine Learning Research (JMLR), **6** (2005) 1705.
- [13] S. Abe and G. B. Bagci, Phys. Rev. E, **71** (2005) 016139.
- [14] E. M. F. Curado and C. Tsallis, Journal of Physics A, **24** (1991) L69.
- [15] C. Tsallis, R.S. Mendes, and A. R. Plastino, Physica A, **261** (1998) 534.
- [16] S. Martínez, F. Nicolás, F. Pennini, and A. Plastino, Physica A, **286** (2000) 489.
- [17] S. Abe, Phys. Rev. E, **79**, 4 (2009) 041116.
- [18] S. Abe, Europhys. Lett., **84**, 6 (2008) 60006.
- [19] H. Hasegawa, Phys. Rev. E, **82**, (2010) 031138.

- [20] H. Hasegawa, J. Math. Phys., **51**, (2010) 093301.
- [21] R. C. Venkatesan and A. Plastino, Physica A, **389** (2010) 1159.
- [22] E. P. Borges, Physica A, **340** (2004) 95.
- [23] J. Naudts, Physica A, **340** (2004) 32.
- [24] W. Stummer, Proc. Appl. Math. Mech., **7** (2007) 1050503.
- [25] W. Stummer and I. Vajda, "On Bregman Distances and Divergences of Probability Measures", 2009. arXiv: 0911.2784.
- [26] T. Wada and A. M. Scarfone, Eur. Phys. J. B, **47** (2005) 557.
- [27] S. Furuichi, J. Math. Phys., **47** (2007) 023302.
- [28] S. Kullback, Information Theory and Statistics, Wiley, New York, 1959.
- [29] S. Kullback and M. A. Khairat, Ann. Math. Statist., **37** (1966) 279.
- [30] G. Kaniadakis, M. Lissia, and A. M. Scarfone, Phys. Rev. E, **71**, 048128 (2005).
- [31] I. Csiszár and G. Tusnády, Statistics and Decisions, **1**, 205 (1984).
- [32] Y. Censor and S. Zenios, Parallel Optimization: Theory, Algorithms, and Applications, Oxford University Press (Oxford, 1998).
- [33] G. L. Ferri, S. Martinez, and A. Plastino, J. Stat. Mech.: Theory and Experiment, **2005(04)** (2005) P04009.
- [34] S. Abe, S. Martinez, F. Pennini, and A. Plastino, Phys. Lett. A, **281** (2001) 126.
- [35] S. Abe, Physica A, **300** (2001) 417.
- [36] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. Modha, J. Machine Learning Research (JMLR), **8** (2007) 1919.
- [37] A. Dukkipati, M. Narasimha Murty, and S. Bhatnagar, Physica A, **361** (2006) 124.

Appendix A: Derivation of expression for the canonical probability which maximizes the dual Tsallis entropy

From (26), the maximum dual Tsallis entropy Lagrangian is

$$\Phi_{q^*}[p, \alpha, \beta] = - \sum_i p_i \ln_{q^*} p_i - \alpha \left(\sum_i p_i - 1 \right) - \beta \left(\sum_i p_i E - U \right). \quad (\text{A.1})$$

Employing the stationarity condition: $\frac{\delta \Phi_{q^*}[p, \alpha, \beta]}{\delta p_i} = 0$ for each p_i and the normalization condition: $\sum_i p_i = 1$, yields

$$\begin{aligned} -\frac{(2-q^*)}{(1-q^*)} p_i^{1-q^*} - \beta E - \alpha &= 0 \\ \Rightarrow p_i &= \left[\frac{(1-q^*)}{(q^*-2)} (\alpha + \beta E) \right]^{\frac{1}{1-q^*}}. \end{aligned} \quad (\text{A.2})$$

Employing the Ferri-Martinez-Plastino methodology [33], the normalization Lagrange multiplier α is obtained as follows. Multiplying the first equation in (A.2) by p_i and summing over all indices i yields

$$\alpha = \left(\frac{(q^* - 2)}{(1 - q^*)} \sum_i p_i^{2-q^*} - \beta U \right). \quad (\text{A.3})$$

Substituting (A.3) into (A.2) yields

$$\begin{aligned} p_i &= \left[\left(\sum_i p_i^{2-q^*} - \frac{(1-q^*)}{(2-q^*)} \beta (E - U) \right) \right]^{\frac{1}{1-q^*}} \\ &= \aleph_{q^*}^{\frac{1}{1-q^*}} \left[\left(1 - \frac{(1-q^*)}{(2-q^*)} \frac{\beta}{\aleph_{q^*}} (E - U) \right) \right]^{\frac{1}{1-q^*}}; \aleph_{q^*} = \sum_i p_i^{2-q^*} \\ \Rightarrow p_i &= \frac{\left[\left(1 - \frac{(1-q^*)}{(2-q^*)} \frac{\beta}{\aleph_{q^*}} (E - U) \right) \right]^{\frac{1}{1-q^*}}}{\aleph_{q^*}^{\frac{1}{q^*-1}}}. \end{aligned} \quad (\text{A.4})$$

Thus (27) is derived.